(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2001 (08.03.2001)

PCT

(10) International Publication Number
WO 01/16805 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US00/24257

(22) International Filing Date:
1 September 2000 (01.09.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/152,500	2 September 1999 (02.09.1999)	US
60/153,593	13 September 1999 (13.09.1999)	US
09/430,450	29 October 1999 (29.10.1999)	US

(71) Applicant: CHILDREN'S MEDICAL CENTER CORPORATION [US/US]; 300 Longwood Avenue, Boston, MA 02115 (US).

(72) Inventors: BUTTE, Atul, Janardhan; 11 C Parley Avenue, Jamaica Plain, MA 02130 (US). KOHANE, Isaac, S.; 227 Summit Avenue #W310, Brookline, MA 02446 (US).

(74) Agent: RODRIGUEZ, Michael, A.; Testa, Hurwitz & Thiebeault, LLP, High Street Tower, 125 High Street, Boston, MA 02110 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/16805 A2

(54) Title: A SYSTEM AND METHOD FOR MINING DATA FROM A DATABASE USING RELEVANCE NETWORKS

(57) Abstract: Described are a system and method for mining data in databases to discover significant relationships among variables in the data. An association is established between each pair of variables. From the data, the strength of the each association is calculated. Correlation coefficients can determine the strength of the associations. In another embodiment, the strength of each association is computed according to mutual information. These calculated strengths are evaluated according to a predetermined criterion. All associations that satisfy the criterion are included in one or more relevance networks. Each relevance network is displayed to provide a pictorial view of the relevant relationships among variables in the data.

A SYSTEM AND METHOD FOR MINING DATA FROM A DATABASE USING RELEVANCE NETWORKS

Related Application

This application claims the benefit of U.S. Provisional Application, Serial No. 60/152,500, filed September 2, 1999, and U.S. Provisional Application, Serial No. 60/153,593, filed September 13, 1999, both incorporated by reference herein.

Field of the Invention

The invention relates generally to data processing. More specifically, the invention relates to a system and method for mining data from a data set to identify potentially meaningful relationships among variables in the data set.

Background of the Invention

With data accumulating in databases in ever increasing amounts, the task of extracting useful information from the data, called data mining, has grown into an important industry. Data mining techniques aim to identify significant relationships among variables in the data. In the field of genomics, for example, human genome sequencing and microarray technology have produced vast quantities of data that may hold the secret to identifying the functions of newly discovered genes. One discipline in particular, called bioinformatics, employs various techniques to mine genomic databases containing sequence,

organism, and expression data to identify clusters of genes having related functionality. As discussed below, current techniques using RNA expression data for identifying gene clusters generally fall into three types: those techniques that use simple criteria matching, those that use Euclidean distance, and those that perform comprehensive pair-wise comparisons.

The simple criteria matching technique measures RNA expression levels before and after an intervention. For each gene, fold-differences are calculated. The genes are then sorted according to the calculated fold-differences. Genes showing a fold-change greater than a given threshold are "clustered" with the intervention.

Techniques that use Euclidean distance include self-organizing maps. The self-organizing map technique represents genes as multi-dimensional points in a multi-dimensional space. Coordinates for these points represent expression levels of each gene at various moments in time. A grid of centroids is imposed in the multi-dimensional space, and the centroids are allowed to drift. Each centroid drifts towards a collection of points. When the drifting completes, the centroids identify clusters of genes that exhibit similar time-course behavior. In this way, related genes have a smaller Euclidean distance in the multi-dimensional space. However, large numbers of dimensions can cause the technique to become computationally intensive. Moreover, the resulting gene/ time course clusters provide

little information about specific gene-to-gene relationships among the genes in the clusters.

Techniques that perform comprehensive pair-wise comparisons generally compare each gene against each other gene using a metric. One particular technique creates a vector for each gene. The vector is made up of expression levels taken at various times. Each gene is compared against each other gene by recording the correlation coefficient between the corresponding vectors. The technique then constructs a phylogenetic-type tree with branch lengths between genes being proportional to the correlation coefficients. However, phylogenetic-type trees, in general, do not show more than the most correlated relationships of each gene, omitting the lesser correlated, yet potentially significant relationships.

Another technique combines the Euclidean distance and pair-wise comparison techniques by constructing phylogenetic-type trees with branch length proportional to the Euclidean distance between genes. The coordinates again represent expression levels at various time points. Although this hybrid technique provides an alternative to clustering genes, the above-described limitations of both the Euclidean distance and phylogenetic techniques remain present.

Thus, a need remains for a data mining technique that can uncover the multi-faceted relationships of the various variables

in a data set without encountering the problems and limitations of the aforementioned techniques.

Summary of the Invention

The present invention relates to a system and method for producing a network of related variables. An objective of the invention is to group variables occurring in data extracted from a data source in a manner that makes readily apparent any potentially significant relationships among those variables and consequently motivate hypotheses for targeted research. Another objective is to concurrently examine relationships among large numbers of variables.

In one aspect, the invention features a method that obtains data for a plurality of variables. An association between each pair of variables is established. From the data, a strength of the association between each pair of variables is calculated and evaluated according to a predetermined criterion. A network of variables is produced. The network of variables includes each association having a strength that satisfies the criterion. The variables can represent any type of data (e.g., genomic data, financial information, customer transaction, airline travel information, etc.). The network of variables can be graphically displayed.

In one embodiment, the network of variables is produced by

including each established association irrespective of the strength of that association, and subsequently removing each association from the network of variables that fails to satisfy the criterion. One embodiment includes each variable in the network of variables, and subsequently removes that variable from the network of variables if all associations with that variable fail to satisfy the criterion. Removing a variable from the network of variables can produce a plurality of separate networks of variables.

In another embodiment, the method establishes the criterion as a threshold value for the strength of the association. In one embodiment, each association having a strength above the threshold value satisfies the criterion. The strength of the association between the two variables can be calculated using mutual information between the variables. Other embodiments use a linear regression model (e.g., computing a Pearson correlation coefficient) or a non-linear regression model.

In one embodiment, the threshold value is determined by randomly permuting the data for each pair of variables. A strength of the association between each pair of variables is calculated from the permuted data. The steps of permuting and calculating are repeated a predetermined number of times. The strongest association is determined from the strengths of associations determined using permuted data. The threshold value is set equal to the strongest association.

In another aspect, the invention relates to a system for producing a network of related variables. The system includes memory storing data for the plurality of variables. An associator establishes an association between each pair of variables in the network of variables. A calculator calculates the strength of the association between each pair of variables. An evaluator evaluates the strength of the association between each pair of variables according to a predetermined criterion. A network generator produces a network of variables that includes each association that satisfies the criterion.

In another aspect, the invention relates to a system for determining a strength of association between any two of a plurality of variables. The system includes memory, storing data for two or more variables, and a processor in communication with the memory. The processor executes software that (1) establishes an association between each pair of variables to produce a network of variables, (2) calculates from the data a strength of the association between each pair of variables, (3) evaluates the strength of each association according to a predetermined criterion, (4) produces a network of variables that includes each association, (5) removes each association from the network of variables that fails to satisfy the criterion, and (6) graphically displays the network of variables.

Brief Description of the Drawings

The invention is pointed out with particularity in the appended claims. The advantages of the invention described above, as well as further advantages of the invention, may be better understood by reference to the following description taken in conjunction with the accompanying drawings, in which:

Fig. 1 is a block diagram of an embodiment of an exemplary system for mining data in databases according to the principles of the invention;

Fig. 2A is an embodiment of a table including data from a data source for a plurality of variables;

Fig. 2B is an embodiment of a scatter plot of the data for a pair of variables from the table of Fig. 2A;

Fig. 2C is an embodiment of a scatter plot of the data for another pair of variables from the table of Fig. 2A;

Fig. 3 is a flow chart of an embodiment of exemplary process that produces relevance networks using the associations between variables in a data set according to the principles of the invention;

Fig. 4 is a block diagram of an embodiment of a graphical representation of the associations between each pair of variables in the data set;

Fig. 5 is an embodiment of a variable matrix including examples of strength values for each of associations shown in Fig. 4;

Fig. 6 is an embodiment of a table illustrating an exemplary permutation of the data in the table shown in Fig. 2A;

Fig. 7 is an embodiment of a graph illustrating results from an exemplary process used to determine a threshold value for evaluating the strengths of the associations between each pair of variables;

Figs. 8A, 8B, 8C are embodiments of relevance networks produced by applying different criterion to the variables and links shown in Fig. 4 with the exemplary associated strength values of Fig. 5; and

Fig. 9 is an embodiment of a relevance network produced from actual genomic data.

Detailed Description

The invention provides a method and apparatus for mining data from databases. Fig. 1 shows an exemplary embodiment of system architecture 10 including a computer system 20 in communication with a data source 30. A variety of system architectures can be used to practice the invention. The computer system 20 includes a processor and memory (not shown) programmed to perform data mining that discovers relationships among variables in the data according to the principles of the invention. The processor in one embodiment is a 266 MHz Pentium II™ processor, manufactured by Intel Corporation of Santa Clara, California. One embodiment of the computer system 20 is a Sun

Ultra HPC 5000 server running Solaris, manufactured by Sun Microsystems, Inc. of Palo Alto, California.

The data source 30 in one embodiment is a database system, e.g., ORACLE 8™, or data stored in files on a data storage device, such as a hard disk. To extract data from the data source, the processor of the computer system 20 executes data mining software. Such software is written in any programming language, such as C, C++, etc.

The data in the data source 30 represent measurements of multiple variables for various sample cases. For example, in a medical context, the sample cases in one embodiment are individuals and the measured variables are physical characteristics, such as weight, height, age, gender, race, etc. Similarly, the sample cases in one embodiment pertain to a single patient evaluated at different time intervals. In this embodiment, the patient is subject to particular laboratory tests, such as hemoglobin, hematocrit, and thyroxine measurements taken over a period of time. Here, the measured variables are continuous variables.

As another example, the sample cases are RNA expression measurements and the measured variables are genes. As still another embodiment, the sample cases are corporate institutions for which the measured variables are financial data, such as stock prices, price to earning ratios, etc., acquired over time.

In general, the principles of the invention can be practiced to examine any type of data in search of relationships among various measured variables. The invention can mine data from databases containing customer sales transactions, commercial passenger travel information (e.g., airline), financial data, and data collected by laboratories, research facilities, commercial institutions, finance institutions, etc. An advantage is that the invention can exploit existing electronic databases.

In brief overview, execution of the data mining software causes the computer system 20 to access data in the data source 30. The data mining software associates each variable in the accessed data with every other variable and determines the significance of the association between each pair of variables. Significance can be defined according to a predetermined criterion.

From the determination, the data mining software groups together variables into one or more separate relevance networks. Each relevance network represents a group of related variables; that is, each variable in a relevance network has a significant association (as defined by the criterion) with at least one other variable in that relevance network, and does not have a significant association (as defined by the criterion) with variables in other relevance networks. The data mining software outputs each relevance network for display (e.g., at the

computer system). The displayed output makes it readily apparent that a relationship potentially worthy of targeted research was detected among variables in the data.

Fig. 2A shows an exemplary tabular representation 50 of data in the data source. The measured variables A, B, C, D, and E are represented on the x-axis as columns. The sample cases S1, S2, S3, and S4 on the y-axis are represented as rows. This column and row arrangement is exemplary; the sample cases and variables can appear on either the x- or y-axis and remain within the scope of the invention. In addition, the principles of the invention extend to more sample cases and variables other than those shown in Fig. 2A. The table 50 can be completely, densely, or sparsely populated with data values 52. Fig. 2A shows an exemplary data set of twenty entries wherein the table 50 includes fifteen numerical data values (VAL1 - VAL15). Five entries of the table 50 lack a data value, each denoted by a dashed line.

The data values 52 are used to determine the degree of a relationship between each pair of variables. Each pair of data values 52 appearing in the same row in the table 50 represents a data point 54 in a scatter plot. For example, Fig. 2B shows an embodiment of an exemplary scatter plot of the data points 54 produced by the data for variables D and E. The data points are (VAL9, VAL12), (VAL10, VAL13), and (VAL11, VAL14). Fig. 2C illustrates another embodiment of an exemplary scatter plot of

the data points for variables A and E, where the data points are (VAL1, VAL12), (VAL2, VAL13), and (VAL3, VAL15). Scatter plots can be produced for each pair of variables in like manner.

Fig. 3 shows an exemplary process for finding relationships among the variables A, B, C, D, and E according to the principles of the invention. The process obtains (step 60) a set of data from the data source 30. The data in the data set includes values for various variables for the sample cases. The computer system 20 organizes (step 64) the obtained data in the data set. One exemplary data organization is the tabular representation 50 shown in Fig. 2A. The computer system 20 associates (step 68) each variable with every other variable in the data set. Accordingly, an association exists between each pair of variables in the data set.

From the data set, the computer system 20 calculates (step 72) the strength of each association. Here, strength is an indication of how closely the variables are related. A strong association indicates that the variables are closely related; a weak association indicates a low or no relationship between the variables.

Variables can be related to each other in various ways. For example, variables can be related through physiology, such as serum concentration of bicarbonate is related to the alveolar partial pressure of carbon dioxide. Variables can be related through mathematical formulae, such as neutrophil count and

percentage of neutrophils. Some variables can be directly or indirectly related to each other through other variables. An example of an indirect relationship is how thyrotropin-releasing hormone controls thyroxine level through thyroid stimulating hormone.

Other variables can have a relationship with each other relating to a pathologic condition. An example of such a relationship is a relationship between the erythrocyte sedimentation rate, which is an indicator of inflammation, and alpha-1 antitrypsin, an acute phase protein indicative of an inflammatory disease state. Other variables can be related through synonymy. For example, both somatomedin C and insulin-like growth factor-1 refer to the same molecule. Here, the principles of the invention can recognize when distinct variables represent the same thing, although referred to by different names.

In one embodiment, the computer system 20 constructs (step 74) a graphical network of variables using every association established in step 68. In this network of variables, each variable is linked to every other variable (e.g., see Fig. 4).

The computer system 20 evaluates (step 76) the strength of the association between each pair of variables according to a predetermined criterion. In one embodiment, the criterion can be a threshold value. The computer system 20 removes (step 80) the association between each pair of variables if the strength

of that association fails to satisfy the predetermined criterion. For example, the predetermined criterion can require the strength of the association between each pair of variables to be above the threshold value, or otherwise that pair of variables becomes disassociated. In another embodiment, the predetermined criterion can require the strength of the association for each variable pair to be below the threshold value in order for that association to remain.

The computer system 20 also removes (step 84) each variable that has no associations with other variables remaining after step 80; that is, all associations of that variable fail to satisfy the criterion. The remaining associations and variables form one or more relevance networks. In step 88, each relevance network is displayed at the client system 26.

The removal of associations and variables can divide the network of variables into smaller, separate networks. Each such smaller network is a relevance network because that smaller network represents a group of related variables. Each variable in that smaller network has an association with at least one other variable in that network that satisfies the criterion.

In some instances, the criterion may cause the removal of none, one, or multiple associations without the removal of any variables. In such a case, the relevance network includes all of the variables in the data set.

In another embodiment, shown in Fig. 3 with dashed lines, the computer produces (step 74') the graphical network of variables after the strength of each association is evaluated against the criterion in step 76. In this embodiment, the computer system 20 constructs the network of variables using only those associations that satisfy the criterion. Variables appear in this network of variables if there is at least one association with that variable which satisfies the criterion. Thus, this network of variables is constructed as a relevance network because the network of variables includes only those variables and associations that satisfy the criterion throughout construction of the variable network. No associations or variables need to be removed from this variable network, such as described in connection with steps 80 and 84, to produce a relevance network.

Other embodiments of processes for constructing a relevance network from associations that satisfy the criterion can be used to practice the principles of the invention.

Fig. 4 shows an exemplary embodiment of a network of variables 110 graphically representing the associations initially established between each pair of variables. The associations are represented as links 100 between pairs of variables. Each variable A, B, C, D, and E is shown as a node in the network of variables and shares a link 100 with every other variable. For example, variable A shares a link 100 with

variable B, another link 100 with variable C, another link 100 with variable D, and yet another link 100 with variable E. Each link 100 has an assigned value representing the strength of the association between the pairs of variables.

Fig. 5 shows an exemplary matrix 104 containing examples of strength values 108 assigned to each of the links 100 of Fig. 4. The matrix 104 places the variables A, B, C, D, and E on both the x- and y-axes. Each value 108 in the matrix 104 represents the strength determined for the association between the respective pair of variables. As such, the matrix 104 is symmetric, and those entries in the matrix 104 denoted by X are either duplicative of another entry in the matrix 104 (e.g., entries (A, B) and (B, A)), or tautological (e.g., entry (A, A)). Such entries need not be calculated or stored. The values 108 shown are exemplary and selected only for illustrating the principles of the invention.

A variety of methodologies can be used to calculate strength of the association between each pair of variables. The following described methodologies are exemplary, as the principles of the invention can be practiced using any methodology capable of assessing the quality of relationships between pairs of variables. Such methodologies can make quantitative or qualitative assessments of those relationships.

One methodology is to consider the number of data points that are used to establish an association between a pair of

variables. Associations between variables based on a high number of data points are stronger than those associations based on fewer data points. This methodology for establishing the strength of an association can be used alone or in combination with other methodologies, such as those described below.

Another exemplary methodology computes a correlation coefficient (typically denoted as *r*) between each pair of variables. The technique for computing a correlation coefficient can depend upon the kinds of variables in the data set.

One technique uses a linear regression model to compute a correlation coefficient with a value between -1 and 1. A correlation coefficient of 1 indicates a perfect linear relationship between variables with a positive slope, a correlation coefficient of -1 indicates a perfect linear inverse correlation (i.e., a relationship with a negative slope), and a correlation coefficient of 0 indicates no linear relationship. Use of this correlation coefficient detects positive and negative relationships between two variables.

In one embodiment, the correlation coefficient is Pearson's correlation coefficient. The Pearson correlation coefficient can measure the linear association between variables for which the data have been measured over intervals. In another embodiment, the correlation coefficient is a Spearman Rank correlation coefficient. The Spearman Rank correlation

coefficient can be a more appropriate coefficient than the Pearson correlation coefficient when actual numerical values cannot be assigned to variables, but a rank order is assigned to each sample case of each variable.

For a coefficient that is more indicative of a predictable linear relationship between two variables than r, the square of the correlation coefficient, r^2, (typically referred to as the coefficient of determination) can be used. The value of r^2 ranges between 0 and 1. Because the value of r^2 is the square of the correlation coefficient, the value is always positive with respect to the coefficient and tends to enhance the differences between correlation coefficient values that are highly correlated. That is, a correlation coefficient, r, of 0.5 has a r^2 of 0.25, whereas an r of greater than 0.7 has a r^2 of greater than 0.5.

Another technique for computing a correlation coefficient uses a nonlinear regression model. Other statistical methods of computing correlation coefficients between variables are known in the art and can be used to determine the strength of the associations between pairs of variables.

Another exemplary methodology for determining the strength of the association between a pair of variables computes entropy (H) of the variables and the mutual information between each pair of variables. The entropy of a variable is a measure of the information content in that variable. Mutual information is

a measure of the additional information known about one variable when given another variable, and is useful for variables (e.g., color) that do not have a numerical relationship with other variables.

Entropy for a variable is computed using a histogram model for discrete probabilities. A range of values for the variable is calculated. That range is then subdivided into n sub-ranges. The proportion of measurements in sub-range x_i (or frequency) is denoted as $p(x_i)$. As n approaches infinity, the histogram increasingly models the probability density function for the variable.

Entropy can be calculated using the following equation:

$$H(A) = - \Sigma_{i = 1 \text{ to } n} p(x_i) \log2(p(x_i))$$

where log2 is base 2 logarithm. Higher entropy indicates that the data for that variable are more randomly distributed, and thus has higher information.

Mutual information can be calculated by subtracting the entropy of a first variable (A) given an occurrence of a second variable (B) from the entropy of the first variable (A) as represented by the following equation:

$$MI(A, B) = H(A) - H(A|B).$$

Expressed another way, mutual information can be calculated by subtracting the joint entropy of the two variables from the individual entropy of the two variables.

$$MI(A,B) = H(A) + H(B) - H(A,B).$$

A mutual information of zero means that the joint distribution of values for a pair of variables holds no more information than the variables considered separately. A higher mutual information between two variables indicates that one variable is predictable from the other variable. Consequently, mutual information can be used as a metric between two variables related to their degree of independence.

In a biological context, for example, the computer system 20 can use the above-described equations to compute a mutual information relationship between pairs of genes. The higher the mutual information is between two genes, the greater the strength of the association between those genes (i.e., the more likely those genes have a biological relationship).

As described above, the strength of each association is compared with a criterion. The comparison operates as a filter that removes weakly related or unrelated associations and variables from the network of variables to produce one or more relevance networks. Consequently, the setting of the criterion is determinative as to which variables and associations appear in a relevance network.

In one embodiment, the criterion is a minimum number of data points upon which the strength of each association between variables must be based. Any association based on less than that minimum number of data points fails to satisfy the

criterion and is removed from the network of variables. Such an association is deemed weak because of the paucity of data supporting the association. For example, referring to Fig. 2A, if the minimum number of data points is two, than the associations between variables B and A and between variables B and D fail to satisfy the criterion because both associations are based on one data point only, (VAL5, VAL3) and (VAL4, VAL11), respectively. If instead, the minimum number was set to three data points, then all associations with B would fail to satisfy the criteria, and the process described in Fig. 3 would consequently remove variable B from the network of variables.

In another embodiment, the criterion is a threshold value against which the strength of each association is measured. The threshold value can be set using any technique for the purposes of practicing the invention, such as, for example, trial and error.

Another exemplary technique for setting the threshold value randomly permutes the data for each variable. The manner of permuting the data of each variable is independent of the manner used for each other variable. Fig. 6 shows an exemplary permutation of the data in table 50 shown in Fig. 2A. The permutation of the data creates new data points between variables. For example, the permutation shown in Fig. 6 produces two new data points between variables A and C, namely (VAL2, VAL8) and (VAL1, VAL6), which differ from the original

data points shown in Fig. 2A, namely (VAL2, VAL6) and (VAL3, VAL8).

From the permuted data points, strengths of associations between pairs of variables are calculated. The technique used to calculate the strength of associations for permuted data points is the same as that used for the original data points. Accordingly, if mutual information is used to indicate the strength of associations for the original data points, then mutual information is also used for the permuted data points.

The steps of permuting the data and calculating strengths are repeated a predetermined number of times (e.g., 30). The threshold value is then set to the strongest association obtained from the repeated permutations of the data.

Fig. 7 is a exemplary graph illustrating the results of this process for determining a threshold value as applied to actual data taken from 2,467 genes in *Saccharamomyces cerevisiae*. The results are described in the U.S. provisional patent application, filed September 13, 1999, and given serial number 60/153,593, attorney docket number CMC-008PR1, and incorporated by reference herein. Here, mutual information was calculated between measurements of RNA expression between pairs of the 2,467 genes. The distribution of the mutual information appears as filled circles. Mutual information was also calculated using permuted RNA expression measurements. The average distribution of 30 repeated permutations appears as open

circles. The permutations did not produce any associations having a mutual information value greater than 1.3. Accordingly, the threshold value used to filter associations can be set to 1.3. In this example, any associations produced from the original data points having a mutual information above 1.3 could be considered significant.

Figs. 8A, 8B, and 8C show the resulting relevance networks produced by the process described in Fig. 3. A different criterion is applied to the links 100 representing the associations between variables A, B, C, D, and E shown in Fig. 4, having the exemplary associated strength values shown in Fig. 5. The relevance networks of Figs. 8A, 8B and 8C are the results of applying minimum thresholds of .4, .6, and .7 respectively. Links 100 having a strength value below the threshold are removed, and links 100 greater than or equal to the threshold remain. In these examples, the criterion does not require a minimum number of data points.

Fig. 8A displays a relevance network 120 that includes all of the variables A, B, C, D, E, but fewer associations than those shown in the original network 110 shown in Fig. 4. In particular, all but one association between D and the other variables has been removed. The only remaining association with variable D is with variable E. In Fig. 8B, the remaining association between variables D and E also fails to satisfy the threshold value of .6. Consequently, the resulting relevance

network 122 does not include the variable D because the variable D has no associations with any of the other variables that meet the criterion.

Fig. 8C illustrates how the threshold value of .7 has divided the original network of variables 110 into two smaller, separate relevance networks 125 and 125'. The relevance network 125 includes one link between variables A and E, and the other relevance network 125' includes one link between variables B and C.

The graphical representations of relevance networks shown in Figs. 8A, 8B, and 8C are exemplary. Application of the invention works with large numbers of variables. To graphically represent the relevance networks having large numbers of variables, the computer system 20 can execute graph layout software. An example of such software is the Graph Editor Toolkit, developed by Tom Sawyer Software of Berkeley California.

Fig. 9 is an embodiment of a relevance network 130 produced from actual genome data as described in the U.S. provisional application, serial number 60/153,593. This particular relevance network 130 clustered 143 genes out of a data set of 79 RNA expression measurements of 2,467 genes. The graph layout software isolates two branches of genes 132 and 132' attached to the network 130 by a single association. In Fig. 9, the branches are exploded to show some detail regarding the names of

the associated genes. Such branches of biologically relevant gene clusters identify opportunities for further study.

The present invention is useful in a variety of applications. For example, relevance networks produced for normal cells can be compared to those relevance networks produced for various cancer cells to help identify distinctions and similarities. Similarly, the invention enables comparisons between the relevance networks of various cancers. Another example uses the relevance networks to monitor changes of certain variables throughout the treatment of a patient.

The present invention may be provided as one or more computer-readable programs embodied on or in one or more articles of manufacture. The article of manufacture may be a floppy disk, a hard disk, a CD-ROM, a flash memory card, a PROM, a RAM, a ROM, or a magnetic tape. In general, the computer-readable programs may be implemented in any programming language, LISP, PERL, C, C++, PROLOG, or any byte code language such as JAVA. The software programs may be stored on or in one or more articles of manufacture as object code.

Having described certain embodiments of the invention, it will now become apparent to one of skill in the art that other embodiments incorporating the concepts of the invention may be used. Therefore, the invention should not be limited to certain embodiments, but rather should be limited only by the spirit and scope of the following claims.

Claims

What is claimed is:

1. A method for producing a network of related variables, comprising the steps of:

(a) obtaining data for a plurality of variables;

(b) establishing an association between each pair of variables of the plurality of variables;

(c) calculating from the data a strength of the association between each pair of variables;

(d) evaluating the strength of the association between each pair of variables according to a predetermined criterion; and

(e) producing a network of variables that includes each association if the strength of that association satisfies the criterion.

2. The method of claim 1 wherein producing the network of variables includes the steps of:

including each established association in the network of variables irrespective of the strength of that association; and

removing each established association from the network of variables that fails to satisfy the criterion.

3. The method of claim 1 further comprising the step of:

including each of the plurality of variables in the network of variables; and

removing each variable from the network of variables if all associations with that variable fail to satisfy the criterion.

4. The method of claim 1 wherein the removing the variable from the network of variables produces a plurality of separate networks of variables.

5. The method of claim 1 further comprising the step of establishing the criterion as a threshold value for the strength of the association.

6. The method of claim 1 wherein each association having a strength above the threshold value satisfies the criterion.

7. The method of claim 1 further comprising the steps of:

randomly permuting the data for the plurality of variables;

calculating from the permuted data a strength of the association between each pair of variables;

repeating the steps of permuting and calculating a predetermined number of times;

determining a strongest association from the strengths of associations determined using permuted data; and

setting the threshold value equal to the strongest association.

8. The method of claim 1 further comprising the step of graphically displaying the network of variables.

9. The method of claim 1 wherein the step of calculating the strength of the association between each pair of variables uses a linear regression model.

10. The method of claim 1 wherein the step of calculating the strength of the association between each pair of variables includes computing a Pearson correlation coefficient.

11. The method of claim 1 wherein the step of calculating the strength of the association between each pair of variables uses a non-linear regression model.

12. The method of claim 1 further comprising the steps of:

determining the strength of the association between each pair of variables using mutual information.

13. The method of claim 1 wherein the variables are genes.

14. The method of claim 1 wherein the variables represent financial metrics.

15. A system for producing a network of related variables, comprising:

memory storing data for the plurality of variables;

an associator establishing an association between each pair of variables;

a calculator, in communication with the memory and the associator, calculating from the data a strength of the association between each pair of variables;

an evaluator evaluating the strength of the association between each pair of variables according to a predetermined criterion; and

a network generator producing a network of variables that includes each association that satisfies the criterion.

16. The system of claim 1 further comprising a remover that removes each variable from the network of variables if all associations of that variable fail to satisfy the criterion.

17. The system of claim 1 wherein the evaluator further comprises a criterion setter that establishes the predetermined criterion as a threshold value for the strength of the association.

18. The system of claim 1 further comprising a comparator that compares the strength of each association with the predetermined criterion.

19. The system of claim 1 wherein each association having a strength above the threshold value satisfies the criterion.

20. The system of claim 1 further comprising:

a data permutation device randomly permuting the data for each of the plurality of variables; and wherein the calculator, calculates from the permuted data a strength of the association between each pair of variables, and the criterion setter sets the threshold value to a strongest association from the strengths of associations determined using the permuted data.

21. The system of claim 1 further comprising an output device displaying the network of variables.

22. The system of claim 1 wherein the network of variables includes a plurality of separate networks of variables.

23. The system of claim 1 wherein the calculator applies a linear regression model to the data of each pair of variables to determine the strength of the association between that pair of variables.

24. The system of claim 1 wherein the calculator applies a non-linear regression model to the data of each pair of variables to determine the strength of the association between that pair of variables.

25. The system of claim 1 wherein the calculator computes a mutual information value between each pair of variables to

determine the strength of the association between that pair of variables.

26. A system for determining a strength of association between any two of a plurality of variables, comprising:

memory storing data for two or more variables;

a processor in communication with the memory, the processor executing software that (1) establishes an association between each pair of variables, (2) calculates from the data a strength of the association between each pair of variables, (3) evaluates the strength of each association according to a predetermined criterion; (4) produces a network of variables that includes each association; (5) removes each association from the network of variables that fails to satisfy the criterion; and (6) graphically displays the network of variables.

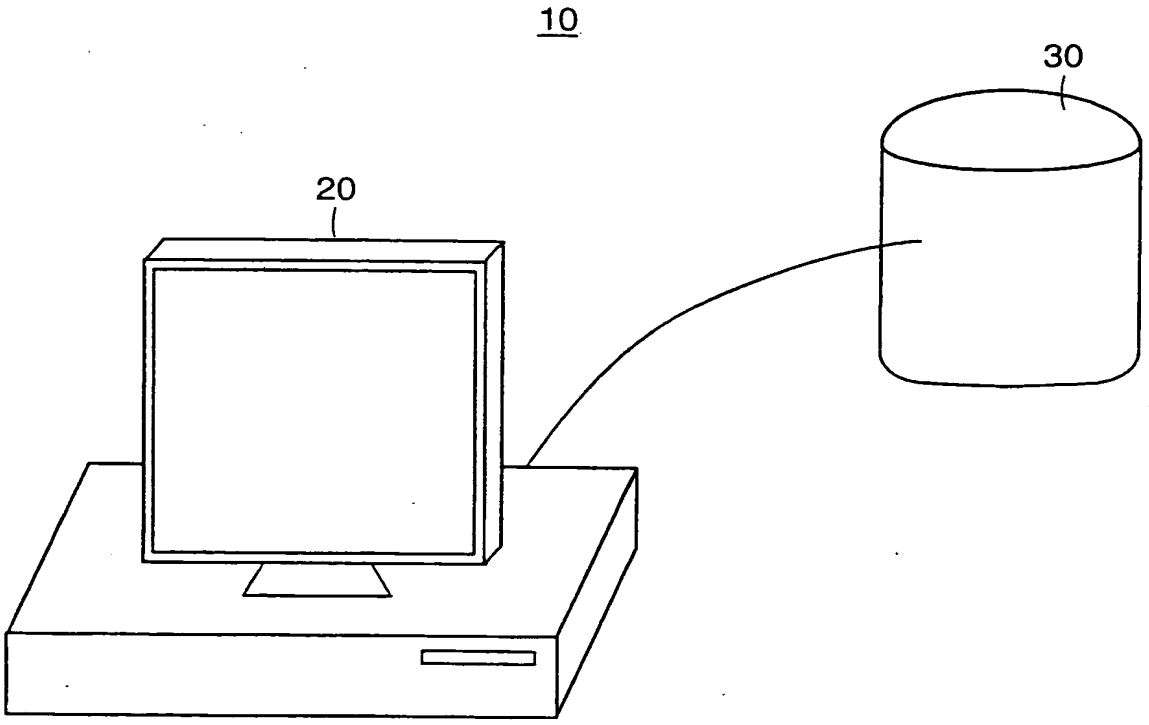

FIG. 1

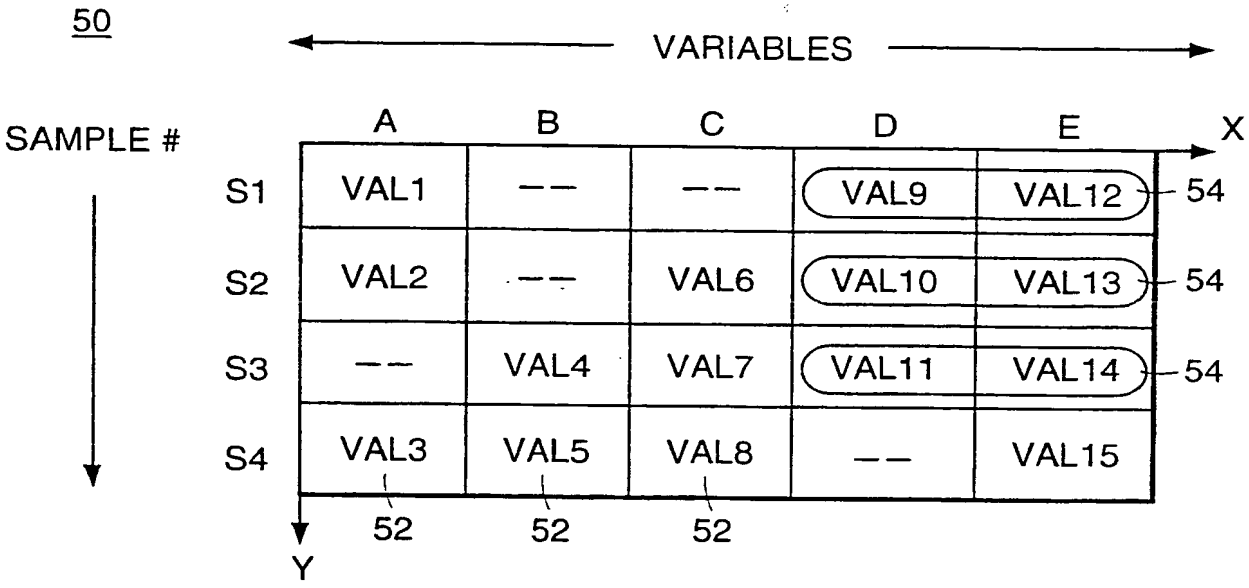

FIG. 2A

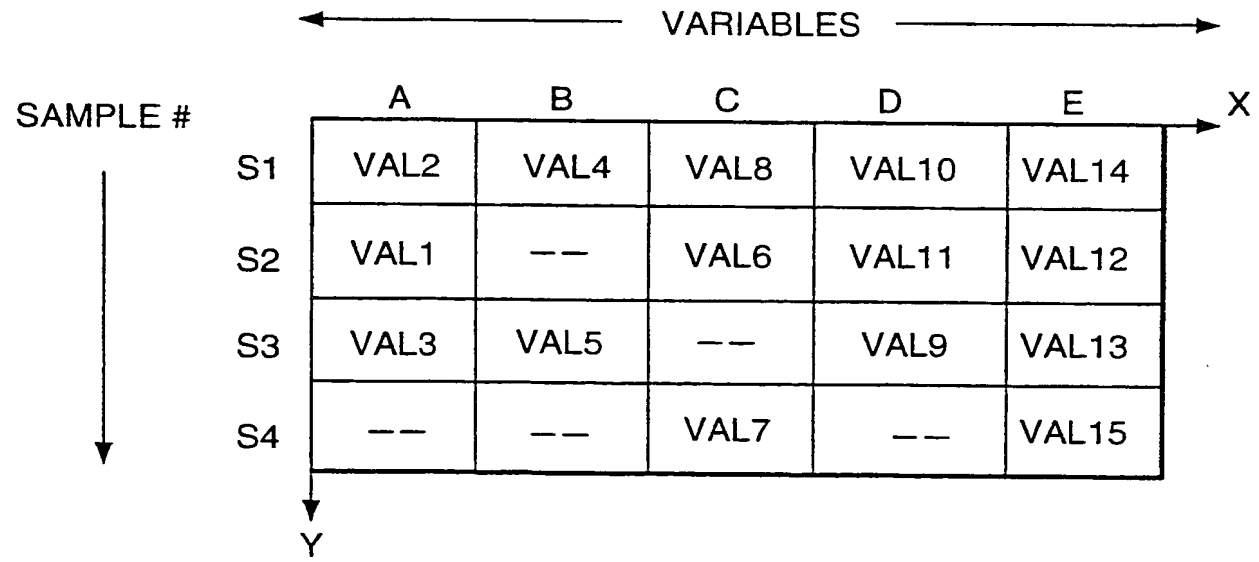

FIG. 6

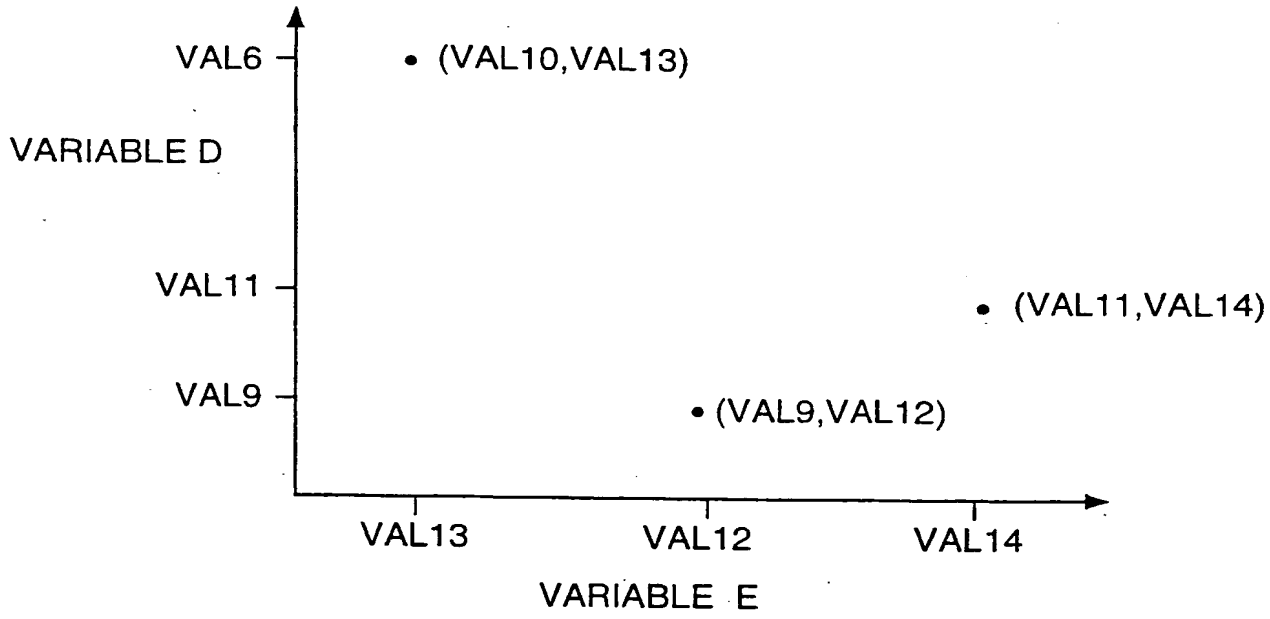

FIG. 2B

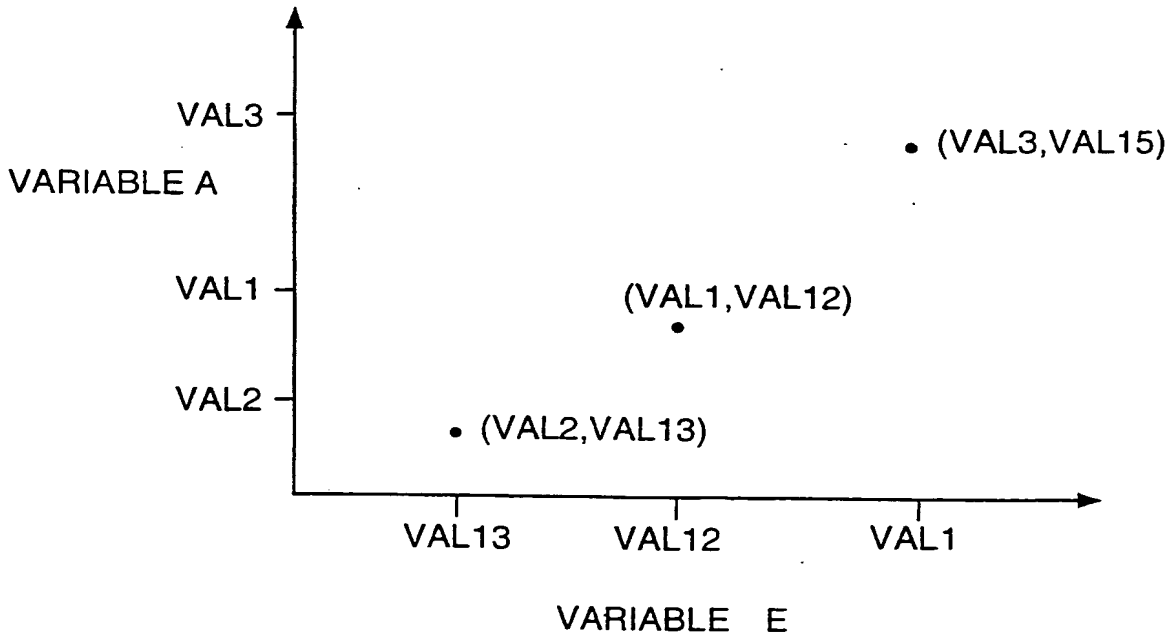

FIG. 2C

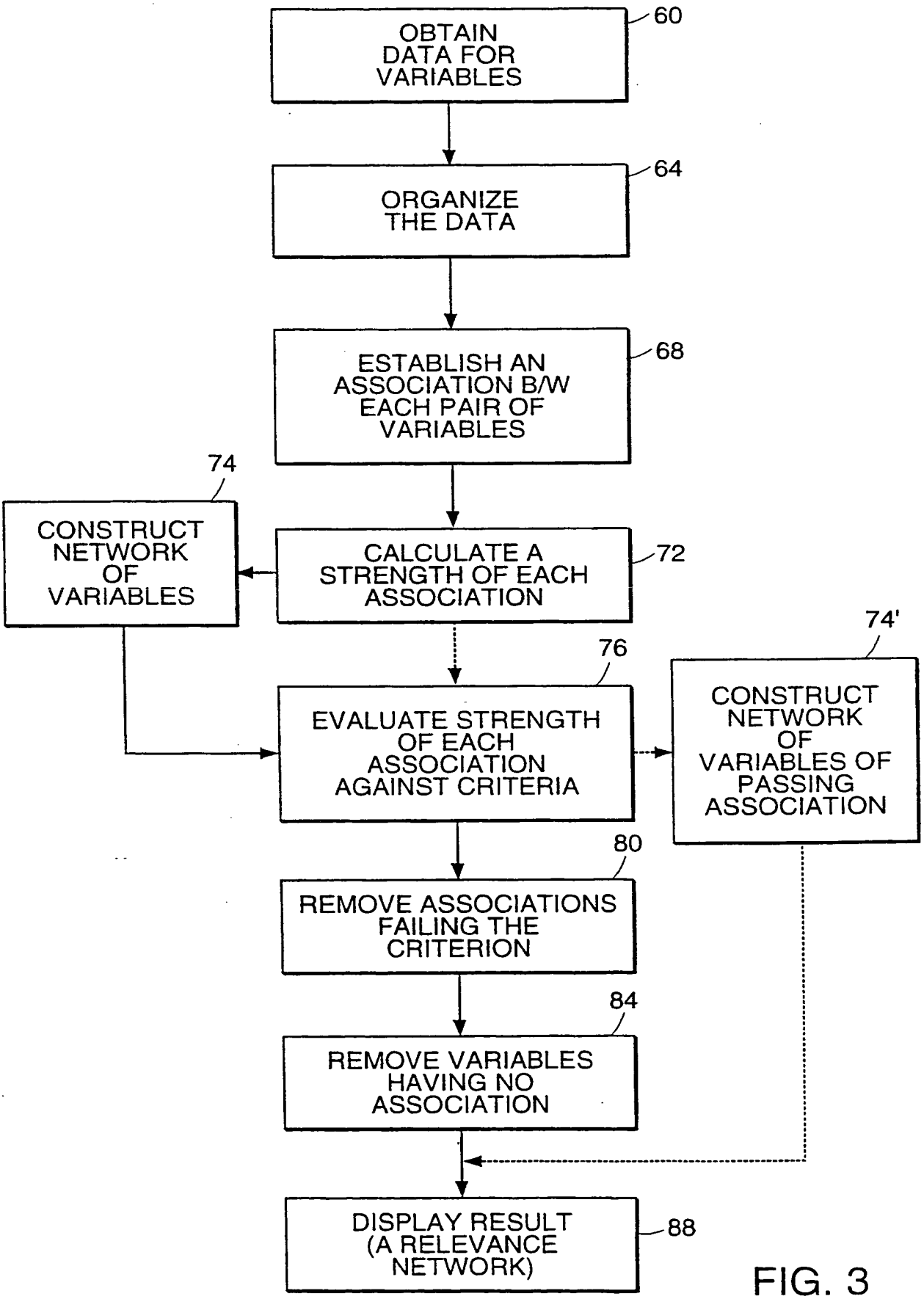

FIG. 3

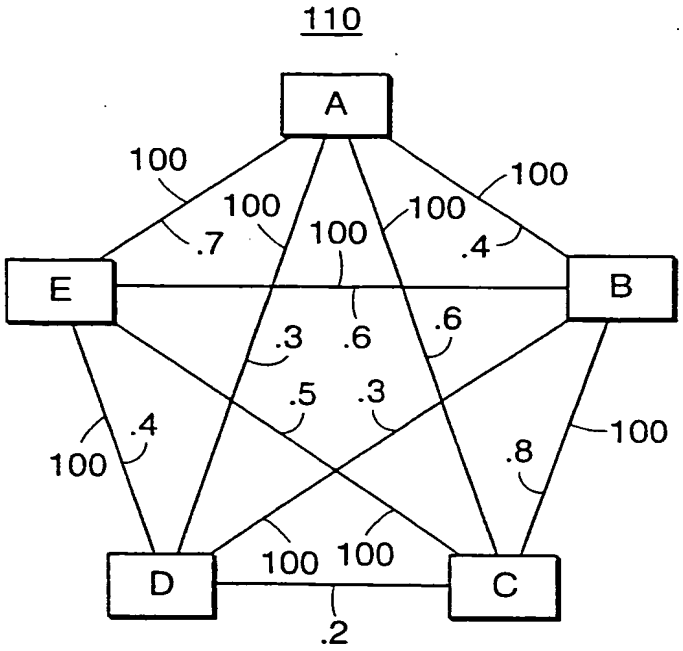

FIG. 4

104

	A	B	C	D	E
A	X	X	X	X	X
B	.4	X	X	X	X
C	.6	.8	X	X	X
D	.3	.3	.2	X	X
E	.7	.6	.5	.4	X

108

FIG. 5

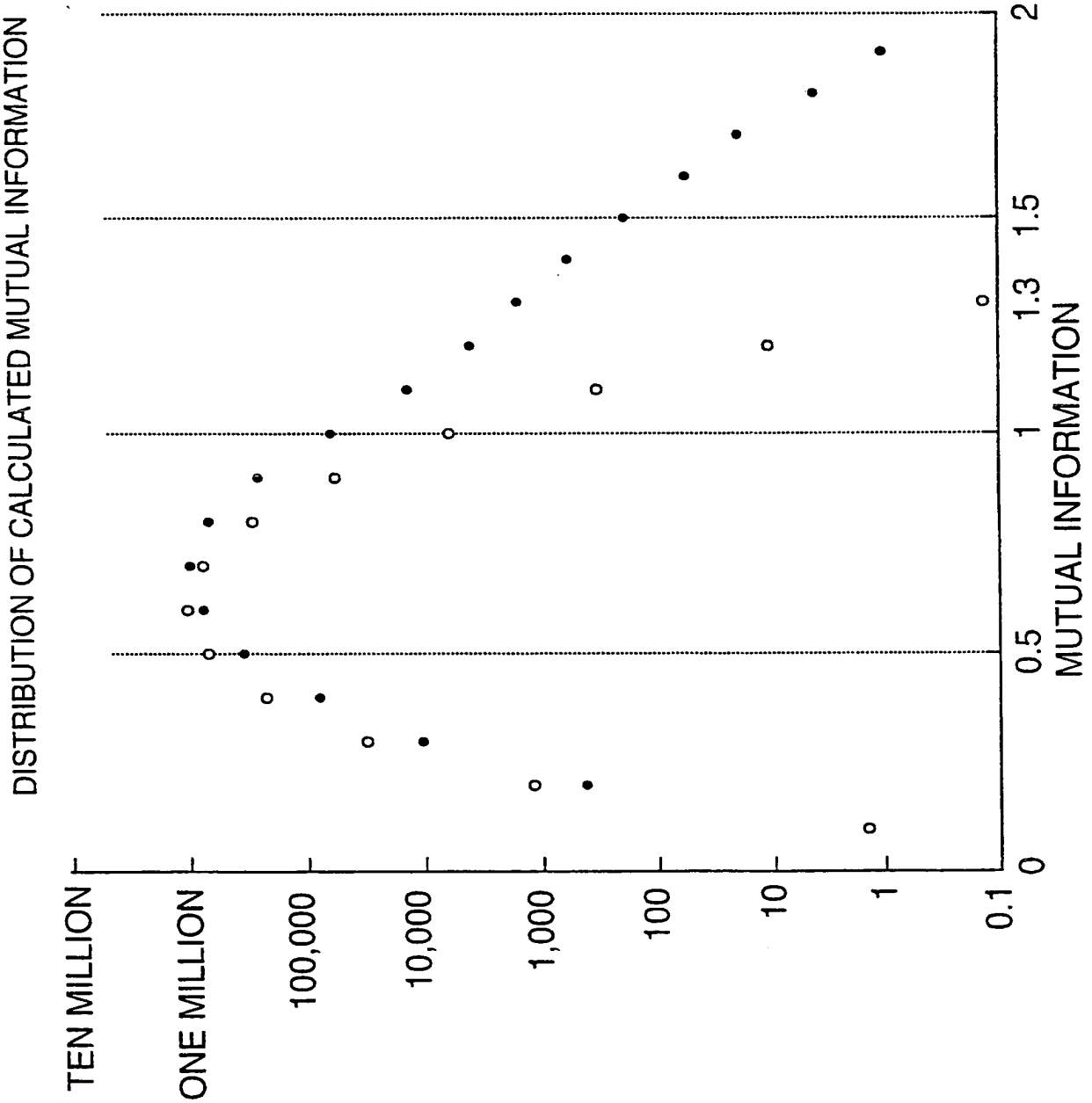

FIG. 7

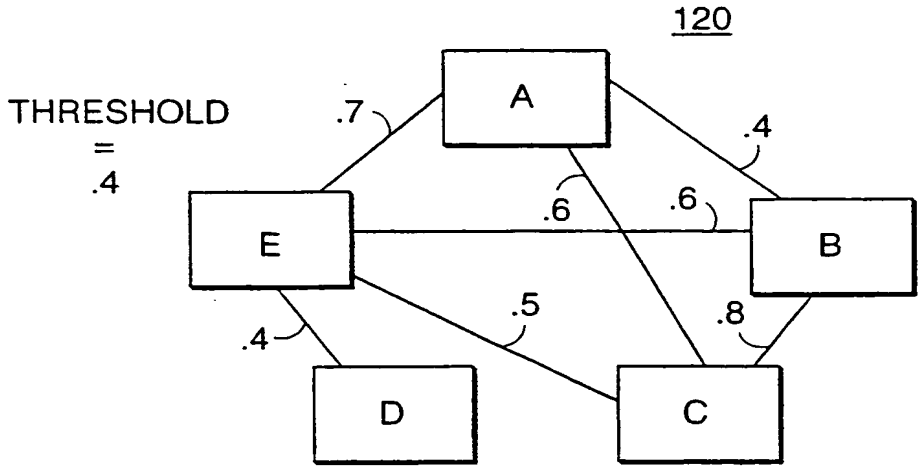

FIG. 8A

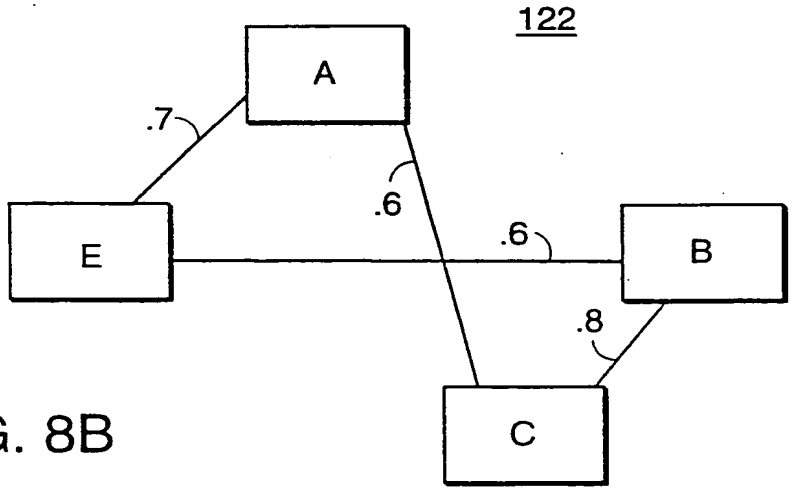

FIG. 8B

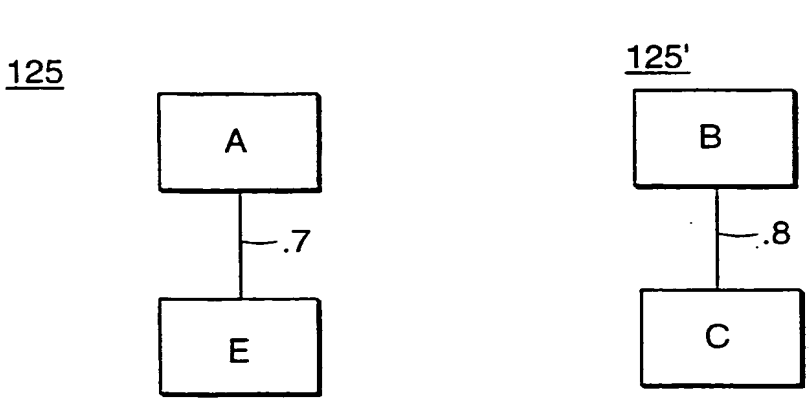

FIG. 8C

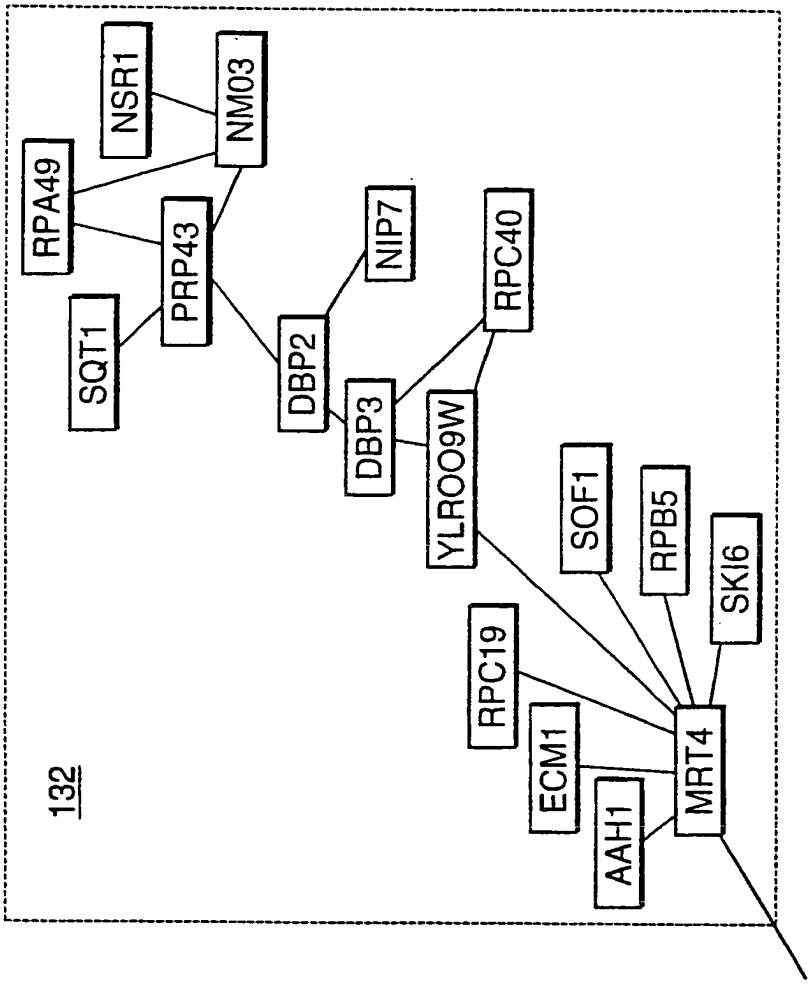

FIG. 9A

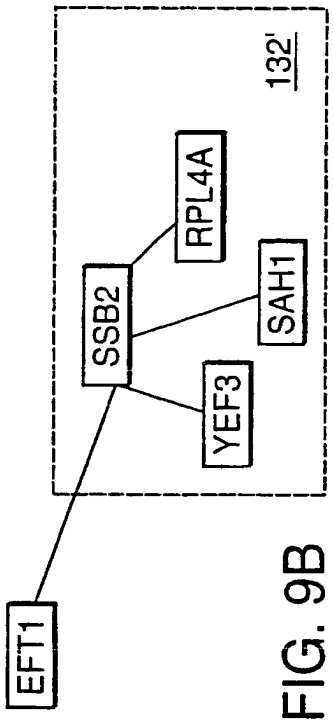

FIG. 9B

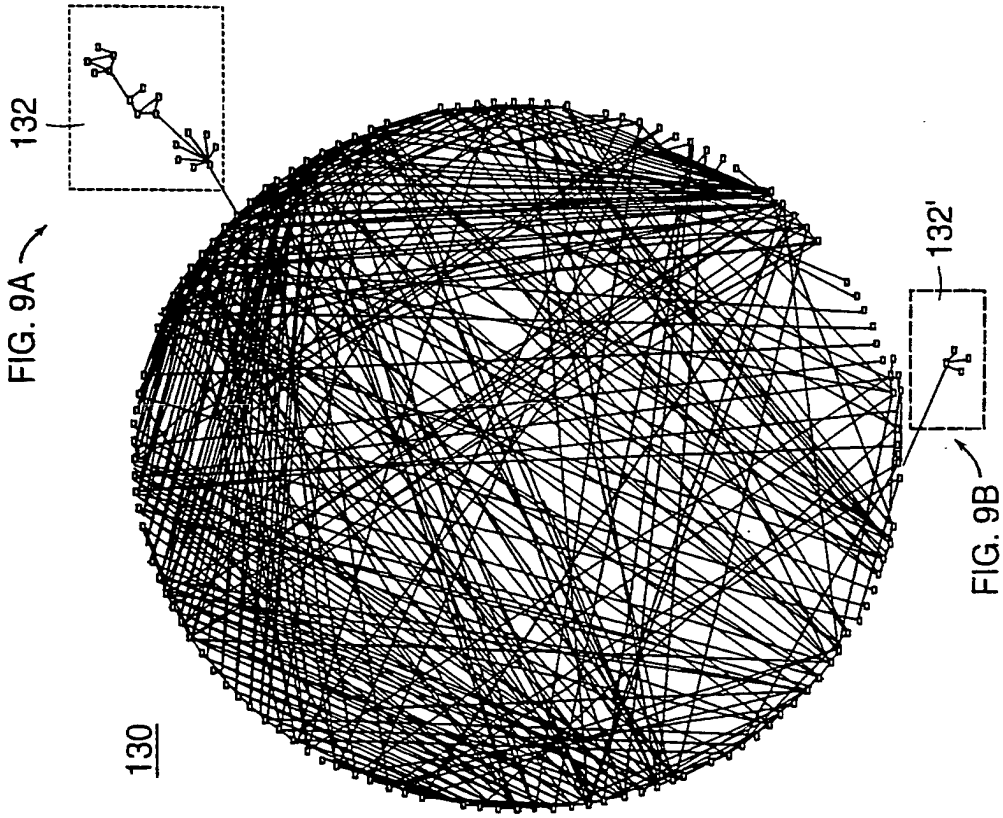

FIG. 9

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2001 (08.03.2001)

PCT

(10) International Publication Number
WO 01/016805 A3

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US00/24257

(22) International Filing Date:
1 September 2000 (01.09.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/152,500	2 September 1999 (02.09.1999)	US
60/153,593	13 September 1999 (13.09.1999)	US
09/430,450	29 October 1999 (29.10.1999)	US

(71) Applicant: **CHILDREN'S MEDICAL CENTER CORPORATION** [US/US]; 300 Longwood Avenue, Boston, MA 02115 (US).

(72) Inventors: **BUTTE, Atul, Janardhan**; 11 C Parley Avenue, Jamaica Plain, MA 02130 (US). **KOHANE, Isaac, S.**; 227 Summit Avenue #W310, Brookline, MA 02446 (US).

(74) Agent: **RODRIGUEZ, Michael, A.**; Testa, Hurwitz & Thiebeault, LLP, High Street Tower, 125 High Street, Boston, MA 02110 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
26 September 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: A SYSTEM AND METHOD FOR MINING DATA FROM A DATABASE USING RELEVANCE NETWORKS

(57) Abstract: Described are a system and method for mining data in databases to discover significant relationships among variables in the data. An association is established between each pair of variables. From the data, the strength of the each association is calculated. Correlation coefficients can determine the strength of the associations. In another embodiment, the strength of each association is computed according to mutual information. These calculated strengths are evaluated according to a predetermined criterion. All associations that satisfy the criterion are included in one or more relevance networks. Each relevance network is displayed to provide a pictorial view of the relevant relationships among variables in the data.

WO 01/016805 A3

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/24257

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CLAVERIE, J.-M.: "Computational methods for the identification of differential and coordinated gene expression" HUMAN MOLECULAR GENETICS, OXFORD UNIVERSITY PRESS, vol. 8, no. 10, 1 September 1999 (1999-09-01), pages 1821-1832, XP002202819	1-6, 8-19, 21-26
A	the whole document	7,20

	-/--	

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 June 2002	17/07/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Jaedicke, M

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	D'HAESELEER, P. ET AL: "Gene Expression Data Analysis and Modeling" PACIFIC SYMPOSIUM ON BIOCOMPUTING 1999 (PSB99), TUTORIAL, 4 - 9 January 1999, pages 1-34, XP002203022 Hawaii, USA	1-6, 8-19, 21-26
A	page 11, last paragraph -page 30, paragraph 1	7,20

A	ALON U ET AL: "BROAD PATTERNS OF GENE EXPRESSION REVEALED BY CLUSTERING ANALYSIS OF TUMOR AND NORMAL COLON TISSUES PROBED BY OLIGONUCLEOTIDE ARRAYS" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, NATIONAL ACADEMY OF SCIENCE. WASHINGTON, US, vol. 96, June 1999 (1999-06), pages 6745-6750, XP002901769 ISSN: 0027-8424 page 6746, right-hand column, paragraph 2; figure 1	1-26

A	EISEN M B ET AL: "Cluster analysis and display of genome-wide expression patterns" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, NATIONAL ACADEMY OF SCIENCE. WASHINGTON, US, vol. 95, December 1998 (1998-12), pages 14863-14868, XP002140966 ISSN: 0027-8424 page 14863, left-hand column, paragraph 1 -page 14864, right-hand column, paragraph 5	1-26

A	BASETT D E ET AL: "GENE EXPRESSION INFORMATICS - IT'S ALL IN YOUR MINE" NATURE GENETICS, NEW YORK, NY, US, vol. 21, no. SUPPL, January 1999 (1999-01), pages 51-55, XP000865988 ISSN: 1061-4036 the whole document	1-26

A	MICHAELS G S ET AL: "CLUSTER ANALYSIS AND DATA VISUALIZATION OF LARGE-SCALE GENE EXPRESSION DATA" PROCEEDINGS OF THE PACIFIC SYMPOSIUM ON BIOCOMPUTING, XX, XX, 1997, pages 42-53, XP000974575 page 45, paragraph 1 -page 46, paragraph 6; figures 1,4	1-26

	-/--	

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	CHEN T ET AL: "Identifying Gene Regulatory Networks from Experimental Data" RECOMB 99, ACM PRESS, April 1999 (1999-04), XP002189969 USA the whole document ---	1-26
A	CHEN Y ET AL: "CLUSTERING ANALYSIS FOR GENE EXPRESSION DATA" PROCEEDINGS OF THE SPIE, SPIE, BELLINGHAM, VA, US, vol. 3602, January 1999 (1999-01), pages 422-428, XP001001103 the whole document -----	1-26